Séparation manuscrit et imprimé dans des documents administratifs complexes par utilisation de SVM et regroupement

Didier Grzejszczak, Yves Rangoni, Abdel Belaïd

Séparation manuscrit et imprimé dans des documents administratifs complexes par utilisation de SVM et regroupement

Didier Grzejszczak, Yves Rangoni, Abdel Belaïd

LORIA
Campus scientifique, BP 239
F-54506 Vandœuvre-lès-Nancy Cedex
{dgrzejsz, rangoni, abelaid}@loria.fr

RÉSUMÉ.
Cet article propose une méthodologie pour la séparation de l'imprimé et du manuscrit dans des images de documents. Les documents à traiter sont majoritairement de type administratif dans un environnement industriel sans contrainte, à savoir une masse quotidienne et importante de pages à traiter avec une grande diversité de contenu et de qualité de numérisation. L'objectif est d'isoler toutes les annotations manuscrites afin d'effectuer par la suite des traitements spécifiques sur le plan du manuscrit et sur le plan de l'imprimé. Nous proposons une solution en plusieurs étapes qui sont: un prétraitement des images, une segmentation du contenu en "pseudo-mots", une reconnaissance par séparateur à vaste marge de la classe d'appartenance, puis une post-correction utilisant le contexte pour affiner la segmentation. Les résultats obtenus sont de l'ordre de 90% de bonne séparation entre l'imprimé, le manuscrit et le bruit.

ABSTRACT.
This paper proposes a methodology for the segmentation of printed and handwritten zones in document images. The documents are mainly of administrative type in an unconstrained industrial framework. We have to deal with a large number each day. They can come from different clients so as to their content, layout and digitization quality vary a lot. The goal is to isolate handwritten notes from the other parts, in order to apply in a second time some dedicated processing on the printed and the handwritten layers. To achieve that, we propose a four step procedure: preprocessing, geometrical layout analysis at pseudo-word level, classification using a SVM, then post-correction with context integration allowing a better quality. The classification rates are around 90% for segmenting printed, handwritten and noisy zones.

MOTS-CLÉS : Analyse de document, segmentation imprimé/manuscrit, SVM, kNN, descripteurs

KEYWORDS: Document image analysis, printed/handwritten segmentation, SVM, kNN, features

1. Introduction

L'objectif du travail présenté dans cet article est le traitement de documents numérisés en vue de séparer plusieurs classes d'information comme celles présentées dans la figure 1. Cette séparation en plan manuscrit / plan imprimé est une étape importante dans le processus de rétroconversion car elle permet très tôt d'éviter des traitements lourds et d'éviter des erreurs lors de la transcription du contenu. En effet le but principal de ce projet est de donner à des outils dédiés chaque plan : l'imprimé passera à l'OCR[1], le manuscrit à l'ICR[2].

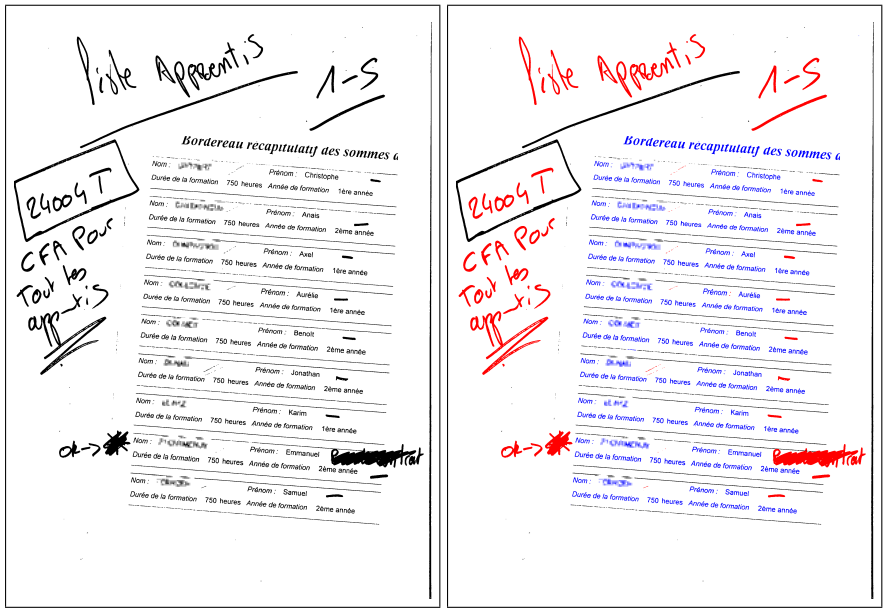

Figure 1. *À gauche, image bitonale d'un document numérisé. À droite, séparation de l'information avec en bleu : texte imprimé ; en rouge : texte manuscrit ; en noir : bruit*

La difficulté majeure de ce problème est la grande diversité des types de document à segmenter, tant au niveau du contenu que de la qualité (Fig. 2). Ces documents sont de type administratif, par exemple des factures, des formulaires, des extraits d'actes de naissance ou des lettres. Ces documents sont rédigés en français et certains font intervenir des annotations manuscrites ou contiennent un mélange d'écritures imprimées et manuscrites par nature. Les documents sont considérés comme aléatoires, étant donné qu'il n'y a pas de contrainte forte sur leur contenu ou sur leur structure. Bien que certains types de documents se répètent, certains sont uniques et doivent quand même être correctement traités. En revanche, si un document est trop difficile à classer ou présente des problèmes, il peut être rejeté.

1. www.abbyy.com
2. www.a2ia.com

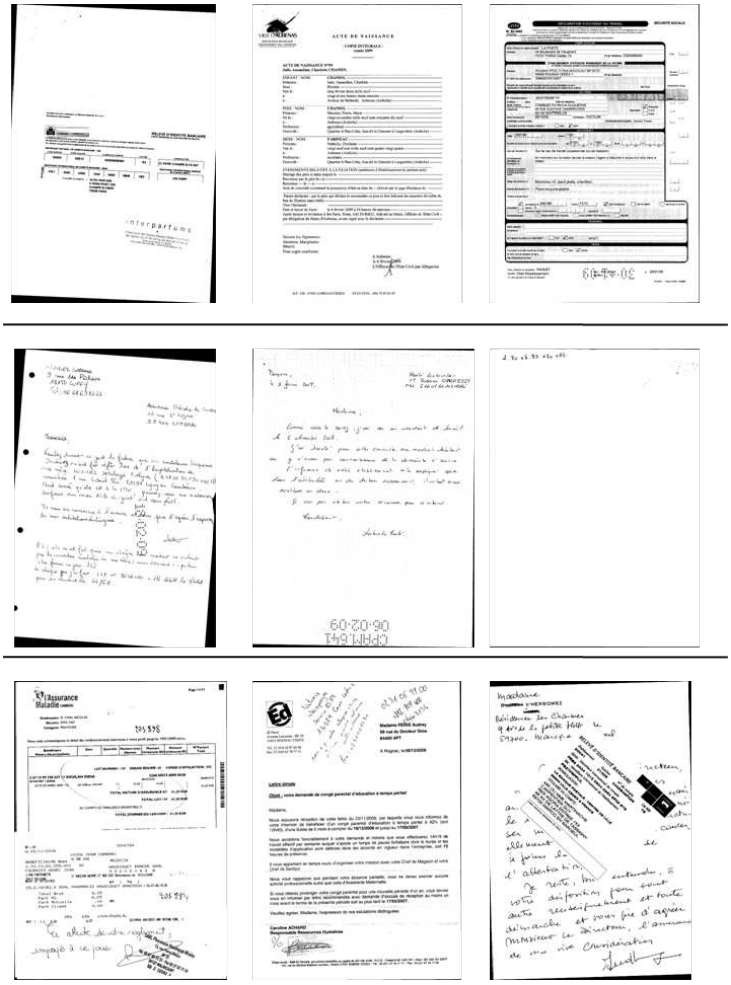

Figure 2. *Exemples de documents de travail*

Les documents de travail ont diverses origines car provenant de plusieurs clients. Il est difficile dans les étapes de prétraitements et d'extraction de caractéristiques de faire des conjectures sur les documents car ils ont été numérisés de manières différentes, tous n'ont pas les mêmes dimensions ni la même qualité, certains sont fortement bruités à cause de la numérisation, d'autres possèdent un angle d'inclinaison très important ou ont subi de fortes altérations. On peut également trouver des éléments graphiques sur les documents comme des photos, des logos, des dessins. Certaines images sont hétérogènes, c'est-à-dire comprennent plusieurs documents sur la même page. Ces documents peuvent être de même type (recto et verso de carte d'identité) ou de types différents (par exemple lettre, chèque et ticket de caisse). Certaines images sont invo-

lontairement composites et font apparaître plusieurs documents suite à une mauvaise numérisation. En revanche, tous les documents sont disponibles dans le même format TIFF multipage et binaire dans lequel les informations de couleur et de niveau de gris sont perdues.

Bien que le thème de la segmentation soit étudié depuis plusieurs décennies (Kang *et al.,* 2009) et que de nombreuses méthodes aient été proposées pour résoudre certains aspects de la séparation imprimé/manuscrit (Casey *et al.,* 1996), la littérature mentionne peu de travaux sur des documents aussi hétérogènes que ceux que nous devons traiter. Nous signalons aussi que les temps de réponse du système doivent être acceptables, et qu'une méthode demandant trop de paramètres empiriques à fixer est fortement rédhibitoire.

Nous nous sommes inspirés des travaux de Kandan et al. (Kandan *et al.,* 2007) qui, pour distinguer le texte imprimé et le texte manuscrit, proposent une méthode de classification en deux classes en utilisant des descripteurs invariants par translation, rotation et changement d'échelle sur les mots de documents peu bruités. La classification par machine à vecteur support (SVM) et par l'algorithme des k plus proches voisins (kNN) sont comparés, puis les auteurs proposent une étape de reclassification en utilisant une triangulation de Delaunay ce qui permet de définir une relation de voisinage sur laquelle est appliquée une règle de majorité.

Zheng et al. ont proposé deux travaux sur la segmentation de documents fortement bruités. Dans le premier travail (Zheng *et al.,* 2002), l'objectif est de déterminer la segmentation la plus adaptée au problème : une comparaison est faite entre les segmentations en mots, lignes et composantes connexes. Le second travail (Zheng *et al.,* 2004) traite de la classification des mots en sélectionnant 31 descripteurs parmi plus d'une centaine. Il introduit également une troisième classe d'éléments pour tenir compte du bruit. Un classifieur de Fisher est utilisé pour étiqueter les blocs segmentés puis un champ de Markov permet une reclassification en tenant compte du contexte de chaque mot.

Le système que nous proposons suit les étapes de traitement suivantes : prétraitement, segmentation en pseudo-mots, qualification des pseudo-mots, classification des pseudo-mots pour terminer par un regroupement des pseudo-mots de même classe. La succession de ces étapes est donnée par le schéma sur la figure 3.

2. Prétraitement

L'étape de prétraitement est cruciale pour la segmentation imprimé/manuscrit et de manière générale conditionne toutes les étapes de rétro-conversion. Cet article considère cette étape comme déjà effectuée. Nous décrivons brièvement ce qui a été effectué pour se prémunir des défauts les plus récurrents. Ces défauts sont :

– l'inclinaison du document numérisé par rapport au document d'origine ;

– des bordures noires autour du document ;

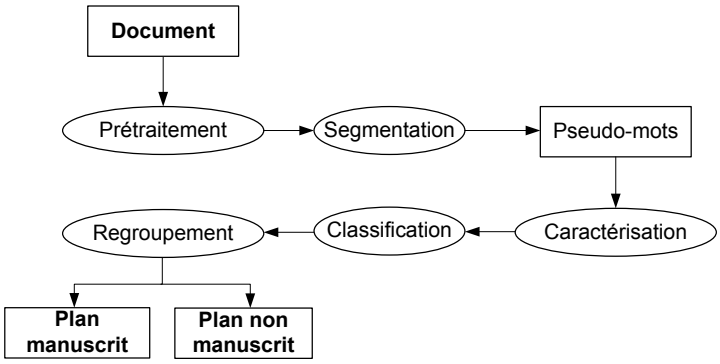

Figure 3. *Vue d'ensemble du système de segmentation*

– du bruit type poivre et sel créé par la numérisation de fond grisé ou à motif sur le document d'origine ;

– des zones ou traits noircis à cause d'impuretés présentes sur la vitre du scanner ;

– du bruit dû à l'apposition de plusieurs documents sur la même page ;

– la présence d'informations non textuelles considérées comme du bruit (logos, cadres, bordures de tableaux, etc.).

Le filtrage est constitué des étapes suivantes :

– suppression des bordures par un système de règles sur la forme et la position des composantes connexes ;

– premier filtrage du bruit par un kfill modifié (Chinnasarn *et al.,* 1998) ;

– détection de l'inclinaison par la méthode RAST (van Beusekom *et al.,* 2010) ;

– second filtrage par un kfill modifié (Chinnasarn *et al.,* 1998) sur le document redressé.

La figure 4 donne un exemple d'application de la procédure de nettoyage.

Après nettoyage de l'image, on exclue les composantes connexes très petites et extrêmement grandes qui ne seront de toute façon ni candidates pour le manuscrit, ni pour l'imprimé.

3. Segmentation en pseudo-mots

Cette étape consiste à créer des zones régulières et stables qui seront utilisées pour étiqueter les parties du document en manuscrit ou imprimé. D'après les conclusions de (Zheng *et al.,* 2002), nous avons opté pour les pseudos-mots. Ils sont plus gros que les composantes connexes, mais plus petits que ce que pourrait donner une segmentation

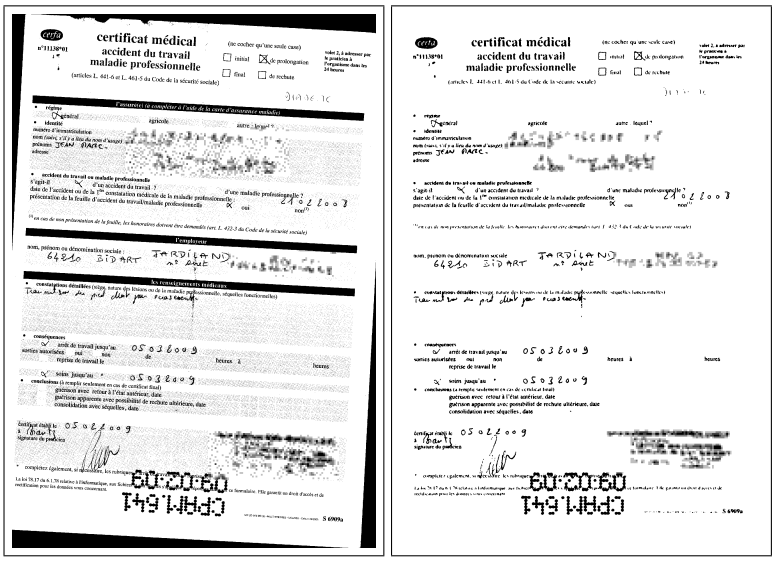

Figure 4. *Exemple d'application de la procédure de nettoyage : à gauche l'image d'origine, à droite l'image nettoyée*

géométrique descendante de la page. Pour de l'imprimé standard et non bruité, le pseudo-mot correspond au mot du texte.

Les documents ayant été nettoyés lors de l'étape précédente, nous avons opté pour une méthode RLSA (Wong *et al.,* 1982), qui donne de bons résultats en des temps extrêmement réduits.

Il s'agit très exactement d'un double RLSA : dans chacune des lignes extraites par un premier smearing, les distances entre les bords des boîtes englobantes des composantes connexes voisines sont calculées. Cela permet de construire un histogramme qui possède généralement une forme caractéristique. Il contient deux pics dominants : le premier correspond à l'écart le plus fréquent (le maximum de l'histogramme) entre composantes connexes, considéré comme la distance entre les caractères d'un même mot. Le second pic correspond au deuxième écart le plus fréquent : celui entre mots de la même ligne. On peut donc appliquer un second RLSA qui permet de segmenter de manière plus fine et d'adapter la segmentation au contenu d'une ligne. La figure 5 illustre la comparaison entre le RLSA original et le double RLSA.

4. Caractérisation des pseudo-mots

La segmentation en pseudo-mots étant effectuée, une étape de caractérisation est nécessaire afin de distinguer la nature de chaque pseudo-mot (imprimé ou manuscrit). Pour ce faire, plusieurs descripteurs ont été étudiés puis sélectionnés. Nous nous

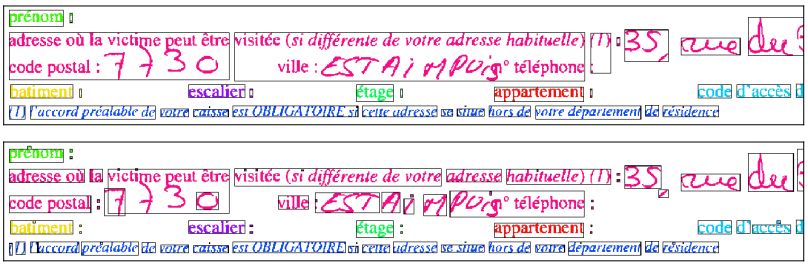

Figure 5. *Comparaison de segmentations : en haut, RLSA simple ; en bas, RLSA double. Les pseudo-mots extraits sont encadrés et les lignes sont identifiées par la couleur des pseudo-mots*

sommes basés principalement sur les travaux de (Zheng *et al.,* 2002), (Zheng *et al.,* 2004), (Kandan *et al.,* 2007) ainsi que (da Silva *et al.,* 2009). Nous avons sélectionné :

– la densité de pixels noirs d'un pseudo-mot ;

– la moyenne et la variance de la largeur/hauteur/aire/densité et ratio des composantes connexes composant le pseudo-mot, ainsi que le ratio de recouvrement des composantes connexes ;

– les moments invariants de Hu ;

– la variance du profil de projection verticale ;

– la distribution verticale des pixels (différence en valeur absolue de la densité entre les moitiés supérieure et inférieure d'un pseudo-mot) ;

– le profil supérieur-inférieur (Kavallieratou *et al.,* 2004) ;

– le dérivée maximale du profil horizontal ;

– la longueur et le nombre de segments horizontaux ;

– le bilevel co-occurence ;

– les $N \times M$ grams, avec $N = M = 2$, 4 distances et 15 motifs donnent à eux seuls 60 descripteurs

L'ensemble des descripteurs forme un vecteur de 137 valeurs. Bien que la littérature conseille une sélection des données, les tests sur le terrain nous ont montré que ni le taux d'erreur ni le temps d'exécution diminuent ; nous avons donc utilisé les descripteurs tels quels, sans modification des algorithmes proposés par les auteurs.

5. Classification

La classification des pseudo-mots est confiée à un SVM (Cortes *et al.,* 1995). S'il était question au départ de séparer uniquement l'imprimé du manuscrit, nous nous sommes orientés vers des SVM multiclasses afin de prendre en compte une troisième

classe qui n'est ni du manuscrit, ni de l'imprimé. Cette troisième classe contient donc à la fois le bruit, mais aussi les cas où manuscrit et imprimé sont inséparables (superposition des écritures). Des différentes méthodes d'apprentissage, nous avons retenu celle de (Weston *et al.,* 1998). La mise en pratique a été effectuée sur la plateforme Weka (Hall *et al.,* 2009) et le classifieur SMO avec l'extension du problème à trois classes par la méthode OnevsOne (Mayoraz *et al.,* 1999).

L'utilisation d'un SVM nécessite un apprentissage. Pour entraîner ce classifieur supervisé, il faut créer des documents de vérité, des documents pour lesquels chaque pseudo-mot a une étiquette correcte. Nous avons donc étiqueté manuellement un petit nombre de documents afin de démarrer l'apprentissage du classifieur. En pratique, dès que la segmentation d'un document n'est pas satisfaisante, le document est corrigé manuellement puis intégré à la base d'apprentissage.

6. Regroupement des pseudo-mots

Si la méthode proposée est relativement rapide et sans utilisation d'a priori ou de modèle, son fonctionnement exclusivement ascendant (bottom-up) le handicape dans des situations difficiles. En effet, s'il est simple pour un humain de faire la distinction entre les deux écritures, c'est parce qu'il traite l'information de manière plus globale, avec intégration de contexte et surtout utilisation de modèle. S'il est exclu pour nous de travailler sur des modèles de document à cause de la forte variabilité des pages à traiter, nous nous sommes concentrés sur l'intégration de contexte.

L'idée est d'étudier pour chaque décision son voisinage et de corriger les erreurs de classification par des méthodes de regroupement. Nous avons expérimenté :

– le regroupement par les plus proches voisins : pour chaque pseudo-mot à vérifier, les k pseudo-mots les plus proches au sens de la distance entre boîtes englobantes sont considérés. Si une classe a été attribuée à plus de la moitié des blocs voisins, alors cette classe est attribuée au pseudo-mot. Pour des résultats tangibles, il convient de fixer une distance maximale au-delà de laquelle on ne considère plus de voisins, et pour optimiser la correction, il est recommandé de donner plus de poids aux voisins sur un axe horizontal afin de regrouper prioritairement les mots d'une même ligne ;

– le regroupement par les plus proches voisins avec contraintes : identique à la méthode précédente avec cette fois-ci une mise à l'écart des petits pseudo-mots (souvent de la ponctuation mal classée) afin de ne pas influencer négativement la correction. Nous avons gardé la règle suivante : l'ensemble des voisins participant au changement totalise un nombre de pixels d'au moins 50% celui du pseudo-mot modifié ;

– regroupement par vote de confiance : la confiance renvoyée par le classifieur aide à la prise de décision. Sur l'idée du regroupement par les plus proches voisins avec contraintes, on examine la confiance du plus proche voisin d'un pseudo-mot considéré. Si cette dernière est plus forte que celle du pseudo-mot, alors il prend la classe du voisin. Au choix, une loi gaussienne ou polynomiale permettent de pondérer la confiance du voisin par son éloignement au pseudo-mot.

7. Expérimentation

L'évaluation de la segmentation imprimé/manuscrit/bruit se fait suivant la mesure proposée par (Shafait *et al.,* 2008). Tous les documents de tests ont donc été étiquetés au pixel près de manière parfaite. Les taux de reconnaissance que nous donnerons seront donc le rapport entre le nombre de pixels correctement étiquetés sur le nombre total de pixels à classifier.

Nous avons utilisé 75 documents pour l'apprentissage du SVM et 24 documents pour le test. La figure 6 donne les différents taux de reconnaissance obtenu par 4 classifieurs dont le SVM (meilleur taux, meilleur écart-type), deux implémentations d'arbre de décision et un perceptron multicouche.

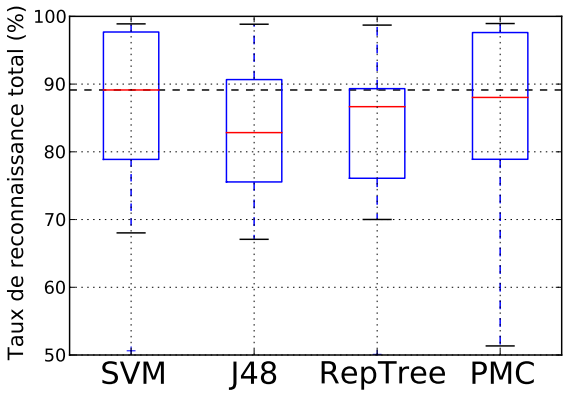

Figure 6. *Évaluation de quatre classifieurs*

Le tableau 7 présente les taux de reconnaissance pour les quatre méthodes de regroupement. Les méthodes kNN utilisent $k = 2$. Les méthodes par confiance utilisent respectivement f_{gauss}, f_{poly2} et f_{poly4} comme fonctions de pondération.

$$f_{gauss}(conf, dist) = conf \times \exp\left(-\frac{10^{-3} * dist^2}{conf^2}\right) \quad [1]$$

$$f_{poly2}(conf, dist) = -5 \cdot 10^{-4}\left(\frac{dist - 1}{conf}\right)^2 + conf \quad [2]$$

$$f_{poly4}(conf, dist) = -10^{-6}\left(\frac{dist - 1}{conf}\right)^4 + conf \quad [3]$$

Nous remarquons que la classification par KNN donne les meilleurs résultats et permet d'améliorer comme prévu les taux de reconnaissance du double RLSA seul. En revanche, les méthodes basées sur la confiance dégradent les résultats. Ceci est principalement dû au fait qu'un seul voisin est pris en compte, ce qui a pour effet de

Taux de reconnaissance	Manuscrit	Imprimé	Bruit	Total
Médiane				
RLSA double	96,1	98,5	**35,7**	89,48
KNN	93,4	98,3	27,3	89,54
KNN avec contraintes	**99,3**	**99,0**	27,9	**90,68**
Confiance gaussienne	94,5	97,7	27,2	87,49
Confiance poly 2 & 4	93,5	97,7	14,2	86,06

Tableau 1. *Évaluation de quatre méthodes de regroupement*

rendre le système peu robuste aux erreurs de classification. Bien qu'il faille encore étudier ce dernier regroupement, nous sommes toutefois dans un certain nombre de situations face à des cas non solutionnables. Il y a bien sûr le cas du manuscrit qui s'enchevêtre avec l'imprimé, et d'autres cas où le regroupement change l'étiquette des annotations manuscrites isolées (un chiffre, un symbole) qui se trouvent noyées dans l'imprimé alors que le classifieur avait donné une bonne réponse. Ce dernier cas nécessite beaucoup plus de contexte et d'interprétation pour être résolu, ce qui n'est pas du ressort de cette méthode rapide de séparation.

8. Conclusion

Dans cet article nous avons proposé une méthodologie pour séparer plusieurs classes d'information dans un document numérisé. L'objectif était de séparer le plan du manuscrit et de l'imprimé, juste après l'étape de prétraitement mais bien avant celle de reconnaissance de contenu dans un système de rétro-conversion de document.

L'application du système devant se faire sur des documents de type administratif et, ce de manière très rapide, nous avons opté pour une série d'extension de méthodes existantes et ayant fait leur preuve sur le terrain.

La méthode consiste à partir d'un double RLSA sur le document permettant d'obtenir facilement les pseudo-mots. Ces derniers servent de "base" à la classification. Des descripteurs sont extraits pour chacun d'eux. Ils ont tous une complexité linéaire ou quasi-linéaire en le nombre de pixels. Les descripteurs sont ensuite envoyés à un SVM multiclasse à noyau gaussien qui s'occupe du premier étiquetage de chaque pseudo-mot. Une seconde analyse est effectuée en étudiant localement le voisinage de chaque pseudo-mot qui peut changer d'étiquette si ses voisins en sont d'une autre. Cette phase d'intégration de contexte permet de corriger un grand nombre d'erreurs. La méthode la plus robuste proposée est le kNN avec contrainte qui, en utilisant un kd-tree, a une complexité totale quasi linéaire en le nombre de pseudo-mots.

Les résultats obtenus sont majoritairement très bons, les taux de reconnaissance avoisinent 90% alors que la base d'apprentissage est très réduite. Aucun rejet n'a été effectué, il pourrait être simple à mettre en œuvre en utilisant le score de confiance

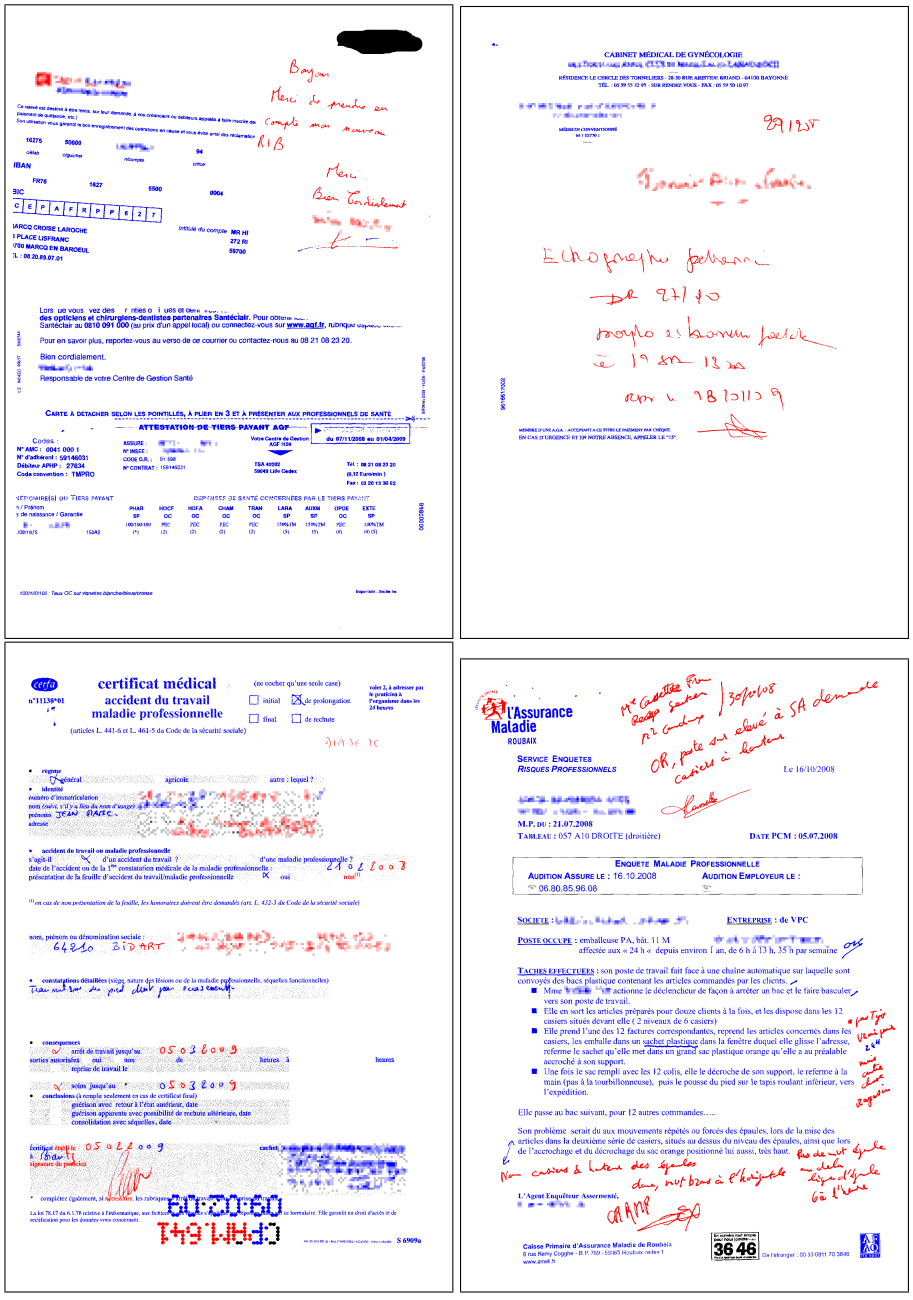

Figure 7. *Exemple de résultats obtenus après le double-RLSA, classification par SVM et regroupement par kNN*

du classifieur et ainsi augmenter encore plus les taux sur les pages ou parties traitées. Nous prévoyons à long terme une approche à apprentissage incrémental, qui consisterait à insérer dans la base d'apprentissage les documents sur lesquels la méthode n'arrive pas à donner des résultats satisfaisants.

9. Bibliographie

Casey R. G., Lecolinet E., « A Survey of Methods and Strategies in Character Segmentation », *Pattern Analysis and Machine Intelligence*, p. 690-706, 1996.

Chinnasarn K., Rangsanseri Y., Thitimajshima P., « Removing Salt-and-Pepper Noise in Text/Graphics Images », *The Asia-Pacific Conference on Circuits and Systems*, p. 459-462, 1998.

Cortes C., Vapnik V., « Support-vector networks », *Machine Learning*, vol. 20, p. 273-297, 1995.

da Silva L. F., Conci A., Sanchez A., « Automatic Discrimination between Printed and Handwritten Text in Documents », *Brazilian Symposium on Computer Graphics and Image Processing*, p. 261-267, 2009.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., « The WEKA Data Mining Software : An Update », *SIGKDD Explorations*, vol. 11, p. 941-954, 2009.

Kandan R., Reddy N. K., Arvind K. R., Ramakrishnan A. G., « A robust two level classification algorithm for text localization in documents », *International conference on Advances in visual computing*, p. 96-105, 2007.

Kang W.-X., Yang Q.-Q., Liang R.-P., « The Comparative Research on Image Segmentation Algorithms », *International Workshop on Education Technology and Computer Science*, p. 703-707, 2009.

Kavallieratou E., Stamatatos S., Antonopoulou H., « Machine-Printed from Handwritten Text Discrimination », *International Workshop on Frontiers in Handwriting Recognition*, p. 312-316, 2004.

Mayoraz E., Alpaydin E., « Support Vector Machines for Multi-class Classification », *International Work-Conference on Artificial Neural Networks*, p. 833-842, 1999.

Shafait F., Keysers D., Breuel T., « Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms », *Pattern Analysis and Machine Intelligence*, vol. 30, p. 941-954, june, 2008.

van Beusekom J., Shafait F., Breuel T., « Combined orientation and skew detection using geometric text-line modeling », *International Journal on Document Analysis and Recognition*, vol. 13, p. 79-92, 2010.

Weston J., Watkins C., « Multi-class Support Vector Machines », *Technical report, Royal Holloway, University of London*, 1998.

Wong K., Casey R., Wahl F., « Document analysis system », *IBM Journal of Research and Development*, p. 647-656, 1982.

Zheng Y., Li H., Doermann D., « The segmentation and identification of handwriting in noisy document images », *Document Analysis System*, p. 95-105, 2002.

Zheng Y., Li H., Doermann D., « Machine Printed Text and Handwriting Identification in Noisy Document Images », *Pattern Analysis Machine Intelligence*, vol. 26, p. 337-353, 2004.